

The Functional Magnetic Resonance Imaging Data Center (fMRIDC): the challenges and rewards of large-scale databasing of neuroimaging studies

John D. Van Horn, Jeffrey S. Grethe, Peter Kostelec, Jeffrey B. Woodward, Javed A. Aslam, Daniela Rus, Daniel Rockmore and Michael S. Gazzaniga*

The fMRIDC, Center for Cognitive Neuroscience, Dartmouth College, 6162 Moore Hall, Hanover, NH 03755, USA

The Functional Magnetic Resonance Imaging Data Center (fMRIDC) (<http://www.fmridc.org>) was established in the Autumn of 1999 with the objective of creating a mechanism by which members of the neuroscientific community may more easily share functional neuroimaging data. Examples in other sciences offer proof of the usefulness and benefit that sharing data provides through encouraging growth and development in those fields. By building a publicly accessible repository of raw data from peer-reviewed studies, the Data Center hopes to create a similarly successful environment for the neurosciences.

In this article, we discuss the continuum of data-sharing efforts and provide an overview of the scientific and practical difficulties inherent in managing various fMRI data-sharing approaches. Next, we detail the organization, design and foundation of the fMRIDC, ranging from its current capabilities to the issues involved in the submitting and requesting of data. We discuss how a publicly accessible database enables other fields to develop relevant tools that can aid in the growth of understanding of cognitive processes. Information retrieval and meta-analytic techniques can be used to search, sort and categorize study information with a view towards subjecting study data to secondary 'meta-' and 'mega-analyses'. In addition, we detail the technical and policy challenges that have had to be addressed in the formation of the Data Center. Among others, these include: human subject confidentiality issues; ensuring investigator's rights; heterogeneous data description and organization; development of search tools; and data transfer issues. We conclude with comments concerning the future of the fMRIDC effort, its role in promoting the sharing of neuroscientific data, and how this may alter the manner in which studies are published.

Keywords: neuroinformatics; functional magnetic resonance imaging; neuroimaging; database; meta-analysis

1. INTRODUCTION

Over the last decade the impact of functional brain imaging on neuroscience has been considerable, advancing scientific understanding of cognitive processes and the neural substrates that underlie them. Neuroimaging techniques have the potential to identify the systems of brain areas responsible for human memory and abstract thinking or to identify elements of the causes for human disorders such as dyslexia and schizophrenia. In recent years, the research potential of functional magnetic resonance imaging (fMRI) has led to a steady and predictable rise in the number of laboratories conducting studies designed to explore the landscape of cognitive function. Still, however, the cost of fMRI experimentation and the infrastructure involved in acquiring the data remains prohibitive for many academic institutions. Indeed, much

of the current body of literature involving neuroimaging research is generated at large, well-funded, medical centres. Scanner time can be costly and the logistics involved in conducting imaging studies of neurological and psychiatric patients can put neuroimaging beyond the means of many researchers. Finally, it is often difficult to design and carry out a rigorously controlled neuroimaging experiment, obtain and analyse the data and scrutinize the results carefully to see if they have implications beyond the scope of the experiment.

After that long and demanding process, it is natural for researchers to be protective of their data. It may, therefore, not be surprising that people feel very passionately about the notion of providing the neuroimaging data from their studies to others. Yet, by its very nature, neuroimaging is a multidisciplinary endeavour, requiring the expertise of physicists, physicians, mathematicians and engineers, among many others. Indeed, successful neuroimaging laboratories tend to be those in which there exists an active and dynamic interaction of these specialities.

*Author for correspondence (michael.s.gazzaniga@dartmouth.edu).

Cognitive neuroscientists alone, although helping to drive the field by collecting the experimental data, cannot be expected to be competent in the details of magnetic resonance (MR) physics, advanced statistical analysis, or even be in a position to interpret all the pertinent information available from these rich datasets. It is for these reasons that data-sharing with peers, both within and between scientific disciplines, is an inherent and necessary component in the science of neuroimaging.

Although, the sharing of brain imaging data is a relatively new idea in the field of neuroscience, in recent years scholars in a variety of other scientific disciplines have begun to realize the usefulness of sharing data. Although just as contentious in their formative stages as neuroscientific data-sharing efforts might be at present (e.g. see Opinion 2000; Kastner 2000; and for comment, Dalton 2000), history has demonstrated that shared databases advance those disciplines that adopt them. Prominent examples of how a database can accelerate scientific progress include the Human Genome Project database (<http://www.ncbi.nlm.nih.gov/>) and the Protein Data Bank (<http://www.rcsb.org/pdb/>). Recent advances in data management and mining now make it possible to use scientific databases as a vehicle for primary research (see for instance, The Sloan Digital Sky Survey, <http://www.sdss.org/>). In this new era, when advances in scientific thinking increasingly require interdisciplinary effort, enabling researchers from many fields to gain access to the raw data from published neuroscientific research is essential for enabling this sort of interdisciplinary scholarship.

(a) *The data-sharing continuum*

The various levels of data sharing in which a researcher may participate can be seen as forming a continuum (figure 1). On one end, 'test' or example data sets may be placed on an anonymous FTP site that other researchers may use for instruction or for early assessment of new image processing algorithms. Then there is the exchange of data with members of one's own laboratory, in which confirmation of results may be the only goal. Where more expertise seems necessary to interpret findings, a researcher may wish to share data with a colleague from another laboratory. Next, descriptions of data from an experiment may be publicly advertised with a view to establishing new collaborations in which novel techniques may be used to analyse and interpret the data (e.g. the peer-to-peer database model, <http://psychology.rutgers.edu/RUMBA>). Finally, after results have been published and are now part of the scientific body of work, researchers may submit their data set to an archive for published data. Thus, other researchers are able to freely access the data that generated the reported findings and on which they may perform their own independent analyses. These new analyses may either confirm the reported results, offer a new interpretation not discussed in the published article, or refute the conclusions of the original authors altogether.

Moving along this continuum the level of scientific quality (in terms of the experimental rigor, control, generalizability, etc.) increases. 'Test' data sets may only be the result of simple, single subject, 'on-off' fMRI experiments, where there is only a rudimentary description of the

experimental protocol and where the data are not intended for use beyond basic analyses. Such data might never be included in a published research report. Conversely, at the other end of the continuum, for data contained in a published study archive, a complete description of the scanner protocol, detailed information about the experimental paradigm, and all the input files needed to reproduce the results reported in the original research article would be expected. Scientific quality, in this case, is implied by the fact that the rationale, statistical analyses and the conclusions drawn by the original researchers have undergone peer-review and are now part of the scientific literature.

The resources required to maintain databases increases with their level of scientific quality. Anonymous FTP sites require little more than the disk space needed to store the compressed representation of the raw data. Such databases are low-cost and need little human supervision. On the other hand, a repository of raw data from published articles would require considerable computer and human resources to maintain. These resources are costly; and necessitate infrastructure considerations, as well as the continued support from funding agencies.

Finally, as data moves from intralaboratory access to interlaboratory and, finally, to full public access, there is an increased need for detailing all aspects and parameters of the data-acquisition procedure. This can impose an increased burden on the experimenter during the data submission process but it is essential for other researchers to be able to completely reproduce effects reported in the published article.

Nevertheless, it is at the more costly and complex end of the continuum we find the greatest potential for advancing neuroscience. For it is here that the scientific record on neuroimaging studies of cognitive processes resides and forms a firm foundation upon which new hypotheses may be generated, scientific discourse can be conducted, and novel methodologies developed and assessed. From this end of the continuum, bridges may be constructed to other neuroscientific databases (e.g. molecular, electroencephalographic, genomic, biobehavioural, etc.) as well as other types of databases, thereby enabling researchers to gather and cross-reference data descriptions and identify convergence of findings. Thus, the involvement of scientists from across disciplines who may scrutinize these data expands the boundaries of what is possible for neuroimaging and, therefore, neuroscience.

2. FOUNDATIONS OF THE FMRI DATA CENTER (fMRIDC) PROJECT

The mission of the fMRIDC (figure 2; <http://www.fmridc.org>) has been to establish a facility for the sharing of functional neuroimaging data within the diverse community of cognitive neuroscience (Kostelec *et al.* 2000). Central to achieving this goal is providing access to fMRI data from peer-reviewed journals that scientists and lay persons alike can use in order to develop and evaluate methods, confirm hypotheses and perform additional analyses. To date, the Data Center is receiving study data from all fMRI analyses published in the *Journal of Cognitive Neuroscience*. This journal's policy of requiring authors to submit their raw study data to the

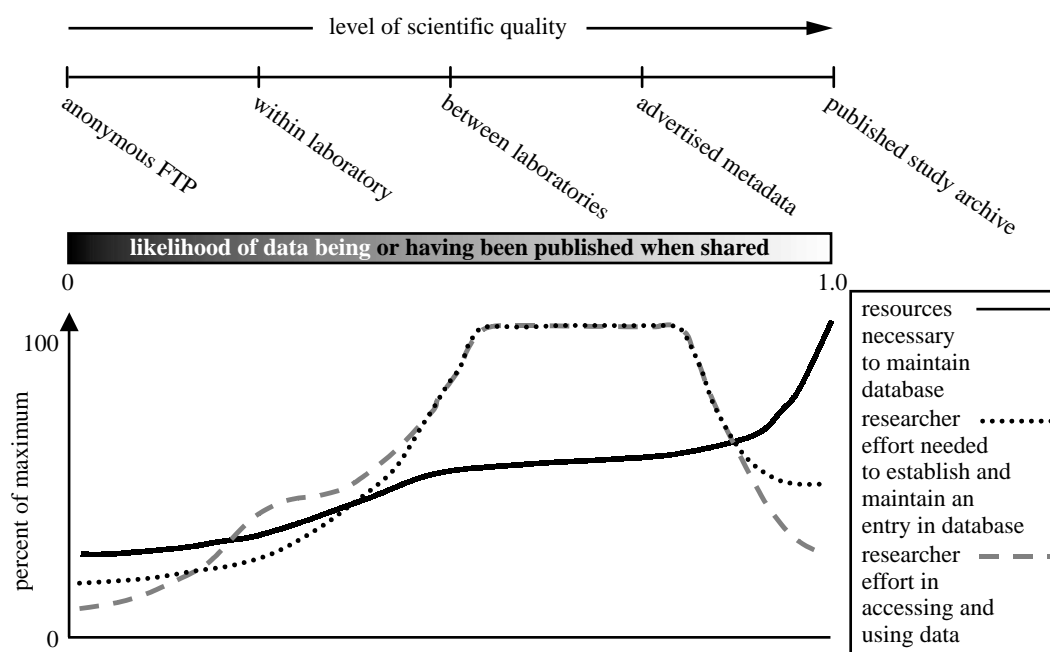


Figure 1. A schematic representation of the hypothesized continuum of data-sharing efforts. The level of loosely defined scientific quality increases as the complexity of the data-sharing effort increases, as does the likelihood of the data being published when it is included in the database. This 'quality' can reflect the experimental rigour, sophistication, sample size and peer-review, among other indices reflective of study control and potential impact to the field. The resources required to store and maintain the database also increase as database complexity increases, being maximal with an archive of raw data from published studies. Researcher effort to contribute and maintain an entry in a database is relatively low in an anonymous FTP-style database and may be maximal when different laboratories are hoping to exchange meaningful experimental data. This might also be true for databases advertising metadata descriptions of data stored at a secondary location. The effort of users hoping to access the data is also minimal with an FTP archive and likely to be maximal in between-laboratory and advertised metadata databases. Ideally, researcher effort in both submitting and accessing data in published study repositories is considerably less than in between-laboratory and advertised metadata approaches. Yet, the greatest scientific value and usefulness for primary research exists on the right-hand side of the continuum.

Data Center was initiated with a special issue (*Journal of Cognitive Neuroscience* 2000, 12(Suppl. 2)), and the archiving and indexing of the data from this inaugural issue is near completion. In addition, other authors have expressed interest in submitting data from their previously published articles in prominent, peer-reviewed journals.

Currently, users may search the archive using a MEDLINE-inspired query interface. This allows researchers to easily access and search the Data Center's resources since many researchers are familiar with the MEDLINE format. Users can search by author, author-supplied keywords, or words in the abstract, among other variables. An extensible grammar allows for both Boolean operators and field tags to limit search results. Basic searches available now can be enhanced for complex queries on the underlying data and metadata. Flexible field tag settings allow the user to indicate how closely their search phrases should match and weights may be assigned to field tags allowing the user to emphasize those search criteria that are important to them. In the future, the Data Center's query interface will be expanded to include even more complex search and categorization schemes as well as to include sophisticated data-mining capabilities (§ 5(b)).

(a) *Architecture of the fMRIDC database*

For the field of neuroimaging to consider itself mature, results from imaging studies should be capable of replica-

tion and consequent critical re-evaluation by independent researchers. Cross-centre neuroimaging studies (e.g. Casey *et al.* 1998) have demonstrated that fMRI results from different laboratories can be reliably reproduced, provided that all necessary data and information are available to the collaborating researchers. This underscores the fact that for a database to be scientifically beneficial it must contain the original data for other scientists to examine. This includes information concerning the study design (e.g. subjects, scanning sessions, scanners and experimental protocols), as well as brain images (e.g. raw reconstructed, preprocessed images and statistical images). In other words, all data necessary to interpret, analyse and replicate an fMRI study should be contained in the database archive so that scientists may universally share this information.

The first consideration in the design of any neuroscientific database management system is the highly heterogeneous and complex nature of the necessary data. Neuroscientists collect data that is made up of many types, from the storage of simple tabular data in both numerical and textual form to more complex data types such as images, videos and time-series data. For example, the fMRIDC needs to manage data collected from various laboratories employing varied experimental paradigms. The methods used to generate the data in these paradigms are probably quite different. For example, a study investigating working memory and saccadic

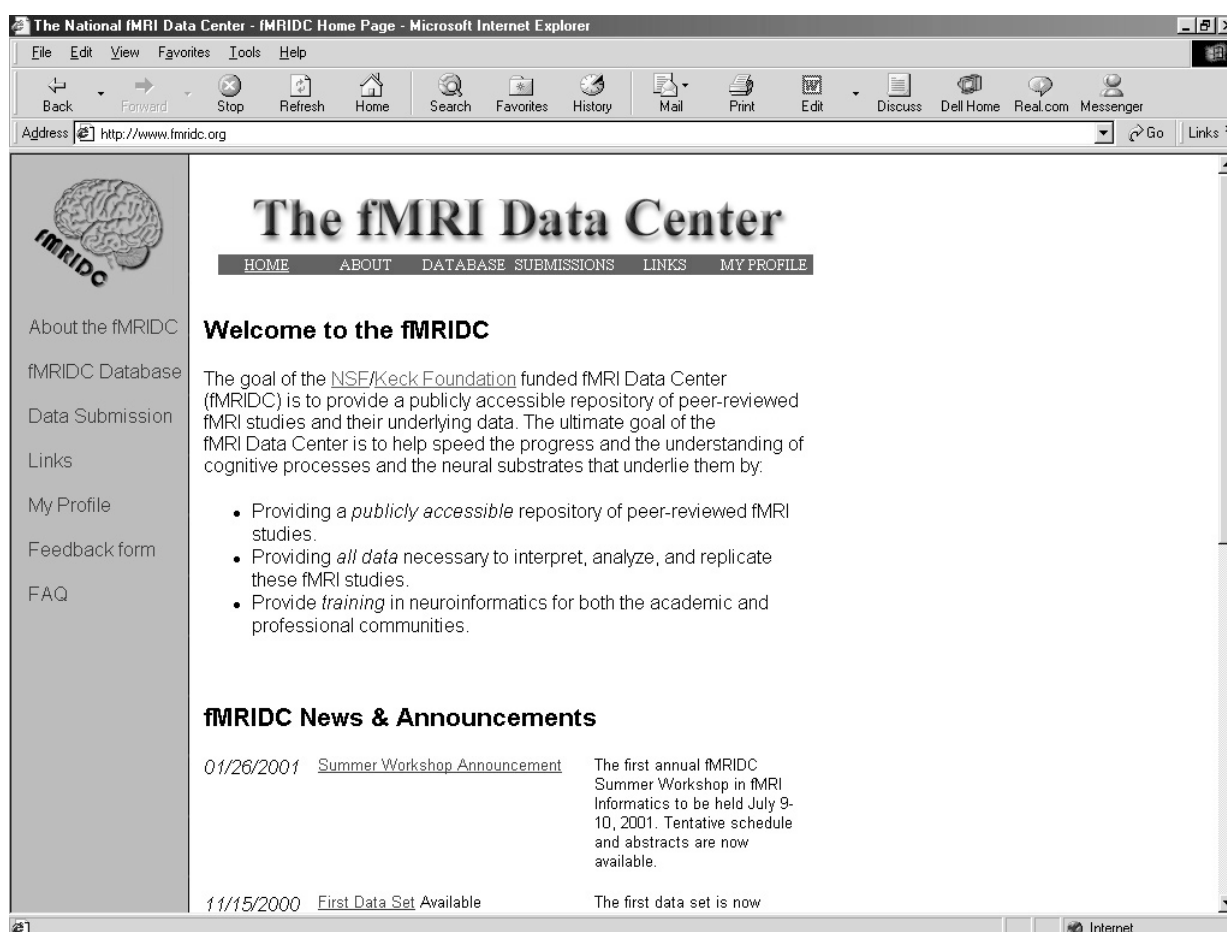


Figure 2. The Data Center hosts a Web site (<http://www.fmridc.org>) that acts as a portal to the database of fMRI studies. To establish a base set of studies in the database, authors from laboratories around the world were invited to submit entire, raw data sets that correspond to the results reported in a special issue of a leading cognitive neuroscientific periodical (*Journal of Cognitive Neuroscience* 2000, 12(Suppl. 2)). All available raw image data from these reports are available for users to order via CD or DAT tape.

behaviour (Postle *et al.* 2000, fMRIDC Accession no. 2-2000-1112R) will have an experimental protocol (e.g. description of the stimuli and their timing) much different from an experiment examining word and pseudoword reading (Mechelli *et al.* 2000, fMRIDC Accession no. 2-2000-11189). In addition to having different experimental protocols, data-acquisition and processing protocols can vary greatly between studies. Therefore, in order to be able to coherently store these datasets we must be able to 'wrap' the experimental data with the information regarding the protocol that was used in collecting the data (i.e. define the MRI scanner-acquisition methodology and experimental design). This 'wrapping' is accomplished by associating all the information regarding the experimental design to the data collected, as well as specifying the MRI scanner-acquisition methodology used to collect the raw image data and the data processing and analysis procedures used to generate the pre-processed and statistical data. In essence, as stated earlier, data provided without the definition of the protocols used in its generation is of little value.

One of the problems in specifying and storing the protocols and data from different studies is their heterogeneity. Each researcher and/or laboratory works with protocols that might be specific to their research focus.

The data they collect and analyse will also contain attributes that are important to their specific research goals. Therefore, in designing the data representation for a scientific experiment a very important principle has to be taken into account: one cannot hope to describe *a priori* all the experimental protocols and research data that neuroscientists will need to incorporate in a rigid database framework. Due to this constraint, any such system designed for the neuroscience community needs to be easily modified to meet a specific researcher's needs without needing major modifications in the system's overall structure.

To address the problems associated with managing the increasingly large and diverse datasets collected throughout the neuroscience community, a novel, extendible object-relational database schema called NEUROCORE was developed as part of the University of Southern California Brain Project (Grethe *et al.* 1996, 1999, 2001; Thompson *et al.* 2001; Arbib *et al.* 1996). By exploiting the object-orientated properties of new object-relational database technologies, NEUROCORE is a database architecture that is completely modifiable while maintaining a standard core structure (figure 3). This methodology can be used to extend the database to contain relevant information concerning the research

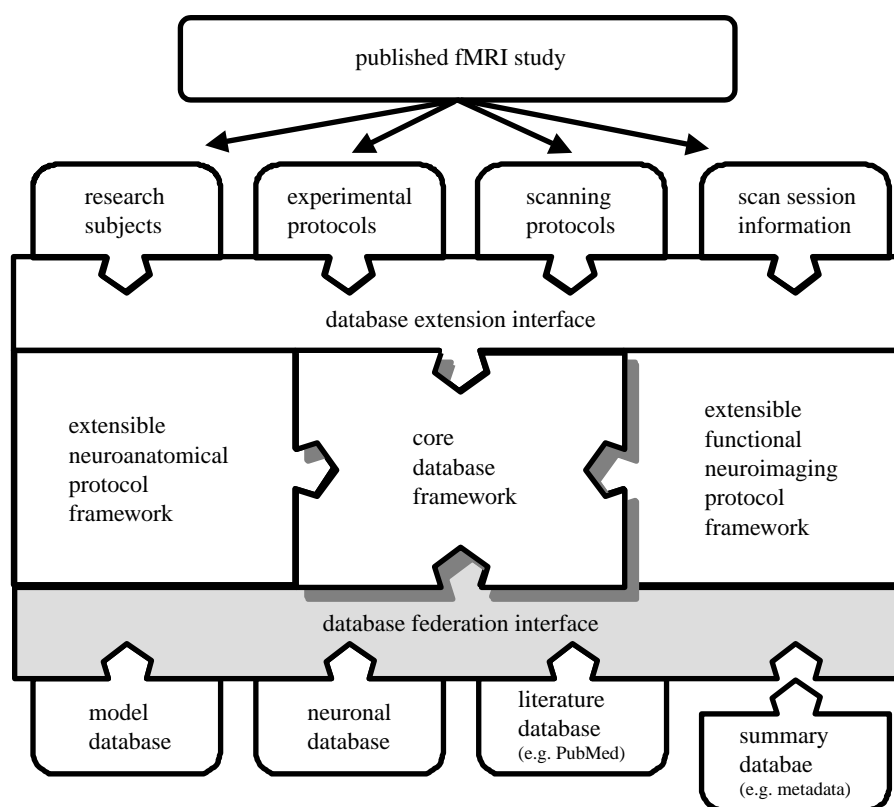


Figure 3. The NeuroCore database architecture permits extensible frameworks for anatomical as well as functional scan information and various scanning parameters that are typical of fMRI studies of cognitive function to be represented. The core database framework employed by the fMRIDC contains a hierarchical description of an experiment. The protocols define the components of an experiment and form the framework for organizing data within this hierarchy. For instance, each experimental protocol consists of a set of manipulations (with stimulus ordering and/or timing information), conditions, groups, etc. These protocols and data are entered into the database by authors through an interface constructed to accept a wide range of information specific to that protocol. In the future, an interface that permits information transfer between other neuroscientific databases will help to enable researchers to search these databases on the basis of metadata retrieved from studies stored at the fMRIDC. (After Arbib & Grethe 2001, p. 17.)

subjects used in an experiment, the experimental data collected, the experimental protocols used and any annotations or statistics normally included with an experiment. The tenets of this framework are as follows:

- (i) *A database is composed of a structured but extensible core.* The core database contains a hierarchical description of the experiment (figure 4). It defines the structure of the experiment and how experimental protocols relate to this hierarchy. This structure allows for the storage of an experiment in a rigorous framework. Each neuroimaging experiment stored in the database is associated with information from each of the subjects' scanning sessions. In turn, each of these scanning sessions is associated with all the scans collected for that subject. All the information in this experiment hierarchy is then also associated with various protocols that define the specifics concerning how the data were collected, the experimental paradigm used to generate the data (figure 5) and the processing steps used to preprocess and analyse the data. The ability to extend experimental descriptions in the database is accomplished through the use of object-relational database technologies (figure 5, bottom). Each descriptor in the database consists of a 'base

tuple' which defines the minimum informational requirements of that descriptor. For example, the base description of a human subject may contain information on a subject's database identifier, age and gender. However, many researchers collect additional data related to the subjects used in their experiments, e.g. subject handedness, diagnostic classification, etc. This basic tuple can be extended to accommodate various experiments. Moreover, these extended tuples can be reused and/or modified for other experiments.

- (ii) *Recognition of the importance of the scientific protocol in storing experimental data.* Many neuroscientific databases only give a cursory description of the protocol for a given data record in the database. In order for data in a database to be useful to all users, it must contain all the specifics regarding the protocol used in the generation of that data.
- (iii) *Account for the complexity of neuroscientific protocols.* Many scientific databases developed currently view scientific protocols as being process-orientated. For instance, in the molecular biology community, an electrophoretic gel is labelled and generates a labelled separation (e.g. see Chen & Markowitz 1995). A more recent example is the US National Institutes of Health (NIH) new Gene Expression

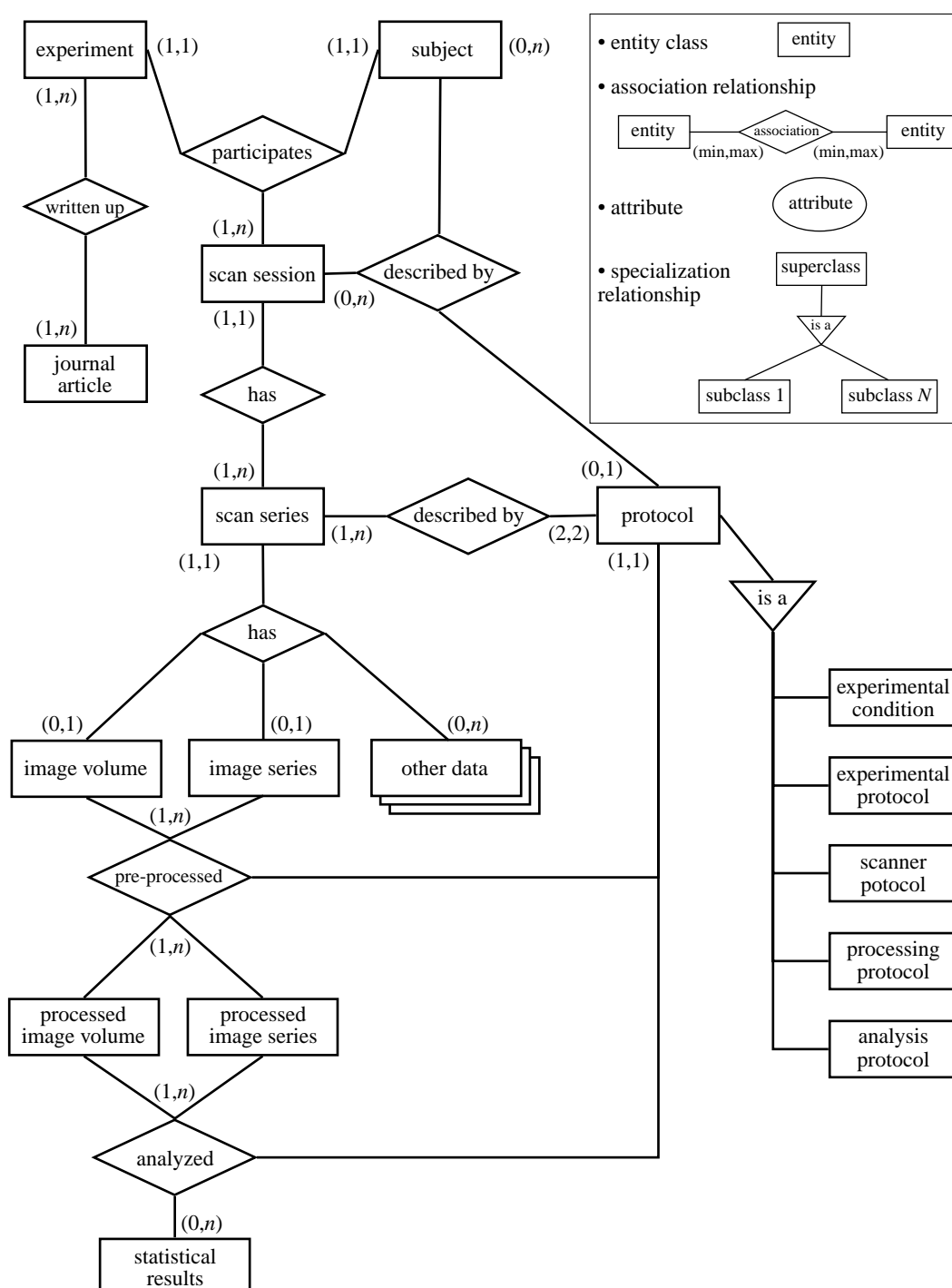
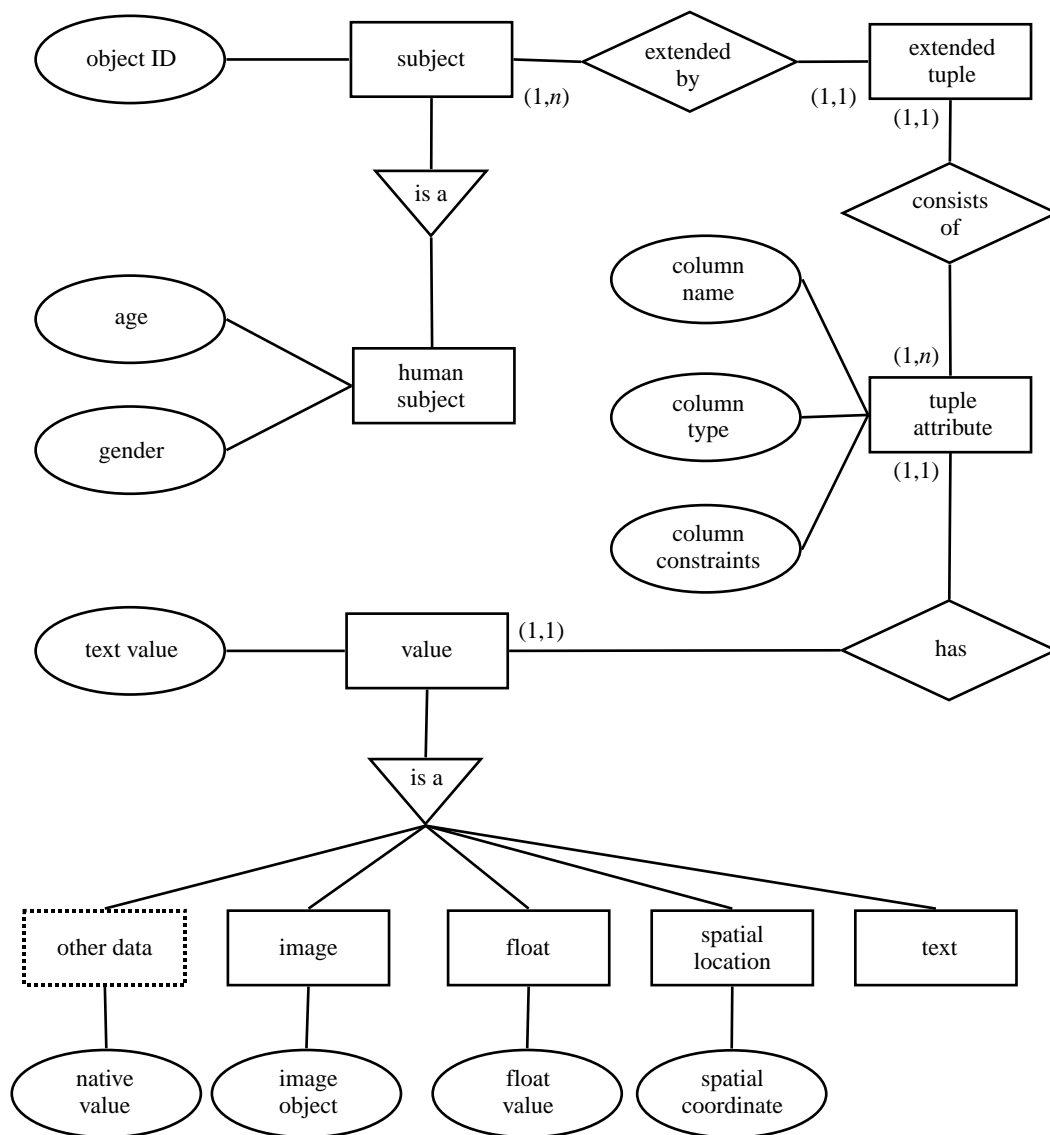


Figure 4. Protocols are associated with many aspects of an experiment, from the generation of the data through to the analysis of the data. The structural aspects of data, including entities, attributes, and relationships are described graphically using an entity-relationship model (Chen 1976). An entity class is an abstraction of a group of objects that share common characteristics and relationships with other objects. An association relationship relates entity classes to one another. Each association also defines the minimum and maximum number of entity classes that must participate in the association relationship (i.e. '(1,n)'). A specialization relationship categorizes a superclass (the parent object) into multiple specialized subclasses (children objects). Through inheritance, each subclass inherits properties (i.e. attributes and relationships) from its superclass. An attribute describes a property of an entity class or a relationship. This diagram shows a schematic of the data-model structure for the experiment and protocol portions of the fMRIDC database. In this case, each imaging experiment involves subjects, who can participate in a number of scanning sessions. Each scan session is made up of a number of scan series (e.g. localizer scan, high-resolution spoiled gradient structural images (SPGR), functional echo planar imaging (EPI) scans, etc.). Each of these scan series can result in the collection of multiple data types (e.g. an EPI scan can result in an EPI image series along with behavioural and physiological recordings).

Omnibus project (<http://www.ncbi.nlm.nih.gov/geo>), a public database designed to receive and make available gene expression data. Expression data is obtained on a variety (and growing number) of

platforms wherein different experiments have different normalizations and complicated experimental descriptions, and may even incorporate a temporal component. Protocols in the neurosciences



	base tuple				extended tuple		
subject attributes	object ID	subject ID	gender	age	handedness	native language	health status
subject no. 3	1014	2-2000-11189-03	male	28	right	English	good
subject no. 4	1015	2-2000-11189-04	female	29	right	English	good
subject no. 3	1016	2-2000-11189-05	male	21	right	English	good

Figure 5. In order to be adaptable, the Data Center needs to be able to store information from varied fMRI experiments and be easily ‘extended’ without the need for major schema modifications. Also, to permit efficient searching of the database, experiments using different protocols and data collection methods need to be stored in the same framework. The ability to extend experimental descriptions in the database is accomplished through the use of object-relational database technologies. The inherent attributes for an entity constitute the ‘base tuple’ that defines the minimum informational requirements for that entity. These base tuples can then be extended with additional attributes through the definition of an extended tuple that defines additional attributes. Furthermore, these extended tuples can be reused and/or modified for other experiments, and be used to guide future interactive data entry forms. The figure depicts the entities and relationships involved in the extension of information regarding research subjects. The table contains an example from a study housed at the fMRIDC (Mechelli *et al.* 2000, Accession no. 2-2000-11189). All subjects in the fMRIDC contain information regarding the subject’s identifier (anonymized), age and gender. However, in the Mechelli study, three additional attributes were defined: handedness, native language and health status.

are considerably more diverse and complicated, consisting of a hierarchy of manipulations that can be ordered as well as specifically timed in relation to one another.

The fMRIDC database model is based largely upon the foundation of NEUROCORE and the lessons learned from it (Grethe *et al.* 1996, 1999, 2001; Grethe & Grafton 1999; Kostelec *et al.* 2000).

(b) *Database storage requirements*

To undertake a project of this magnitude, it is clear that considerable computational resources are necessary (figure 6). The Data Center's main server is a Sun Microsystems (Mountain View, CA, USA) Enterprise 5500 Unix-based platform, which has 2 TB storage capacity. The Data Center also has access to an auxiliary research server with another 2 TB storage, as well as a cluster of Linux workstations and hardware to accommodate a wide variety of digital media (e.g. CDROM, digital linear tape (DLT) and advanced intelligent tape (AIT)). The servers run Informix Internet Foundation 2000 (Informix Software Inc., Westborough, MA, USA) database software and an Apache (Apache Software Foundation, Forest Hills, MD, USA) Web server. In the future it is anticipated that data from imaging studies will be stored offline in a hierarchical fashion (i.e. on disk, CD or DLT). However, the descriptive information for each study (e.g. 'metadata', summary information concerning the data that make up the study as well as detailed information regarding the equipment, imaging protocol and experimental paradigm) will always be stored on disk and will always be immediately available for queries and browsing.

(c) *External advisory board*

For the Data Center to remain focused on the needs of the neuroimaging community and the goals it seeks to attain, a necessary part of the organizational structure includes an External Advisory Board. This Board is made up of leaders in the field of neuroimaging and neuroinformatics with an interest in the sharing and databasing of neuroscientific data (<http://www.fmridc.org/about/people.php>). The responsibility of the Board is to review the Data Center's progress, approve the criteria for depositing data, recommend necessary technical descriptive information for studies, ensure the security measures being undertaken for the data submitted, and provide ongoing oversight of the Data Center's activities.

3. USER INTERACTION WITH THE DATA CENTER ARCHIVE

Researchers exchange data with the Data Center through two simple means: (i) contributing study data to the archive, in which researchers fill out detailed forms describing study protocols and individual subject information, and (ii) requesting study data from the record of studies in the repository. Here we describe these two processes in more detail.

(a) *Contributing study data*

The most important aspect for the storage and organization of any scientific experiment is the storage and description of the experimental protocols involved.

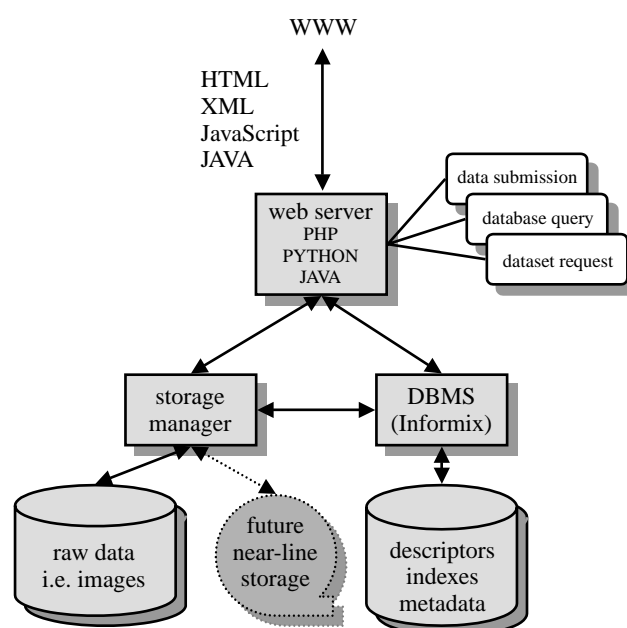


Figure 6. The physical architecture of the fMRIDC computational resource is implemented in three tiers: (i) the client Web browser, using HTML, XML, Java (Sun Microsystems, Mountain View, CA, USA) and JavaScript; (ii) the fMRIDC Web server running PHP (<http://www.php.net>) and PYTHON software (<http://www.python.org>); and (iii) the fMRIDC database management system (under Informix) and storage manager. Currently, users of the database are able to submit their datasets to the Data Center, query the database and request datasets. Future extensions to this interface will allow for detailed viewing of the experimental protocols, enable basic image screening, as well as sophisticated search and categorization tools. Offline tape storage for study image data, which will free up local disk resources, is planned.

Neuroimaging data is associated with a variety of descriptors that define an experiment (e.g. information regarding the equipment, imaging protocol, experimental paradigm and statistical methods). The Data Center requests certain basic information deemed necessary in describing various aspects of the experiment. In addition to this information, the Data Center asks that each submitting author supply any additional information they deem necessary to fully describe their experiment. To accommodate this, the fMRIDC database is designed so that it can be extended to store a researcher's unique descriptors. This framework allows the Data Center to store information for descriptors in a more dynamic fashion and anticipate new descriptors as current methods evolve and new methods emerge. Table 1 shows the basic collection of variables currently requested from authors for describing an fMRI study under this framework.

For submission of brain image data, the fMRIDC accepts all commonly used medical image data file formats used in functional neuroimaging research; in particular Mayo Analyse (Mayo Clinic, Minneapolis, MN, USA), AFNI (Cox 1996) and MINC (Montreal Neurological Institute, Montreal, Canada) formats. For the most part, differences between these file formats is in the manner that the data is organized within the files and the directory structure. We have adopted and developed several tools for file format conversion.

Table 1. fMRI study information collected for the fMRIDC archive

scanner protocol ^a
scanner protocol ID
coil type
pulse sequence type
flip angle (degrees)
echo time (milliseconds)
scanner volume acquisition time (milliseconds)
number of time-points
number of acquisitions
number of dummy scans
number of slices
slice thickness (mm)
slice skip (mm)
interleaved or sequential
slice acquisition
field of view
receiver bandwidth
original acquisition
matrix size (e.g. 256 × 256)
reconstructed image
matrix size (e.g. 256 × 256)
full or partial k-space
orientation of images (radiological or neurological convention)
ramp sampling (yes/no)
echo train length
echo shift if an asymmetric spin echo is used
type of reference scan used for reconstruction
other
subject information
subject ID ^b
experimental group code
gender
age
health status
assessments (e.g. handedness)
medications
other (e.g. diagnostic criterion, etc.)
scan session information
scanner manufacturer
scanner model
scanner software revision
field strength
gradient strength
slew rate
date of scan session
duration of scan session
other
experimental protocol
experimental protocol ID
number of subjects
number of groups
number of functional runs
epoch-related conditions (e.g. timing, stimulus duration, etc.)
event-related conditions (e.g. timing, stimulus duration, etc.)
experimental methods
stimulus regressor files ^c
other (e.g. condition descriptions, associated variables, etc.)

^a Scanner protocol information is targeted for the typical blood oxygen level dependent EPI sequences collected in fMRI studies. Three-dimensional acquisition protocols may not 'fit' with these predefined descriptors. However, the user can extend descriptors so that any acquisition protocol may be stored within the database.

^b As mentioned in the text (§ 4(a)), information contributed that may be used to identify subjects is removed or recoded to comply with US federal guidelines on the protection of human subject data.

^c Experimental stimulus regressor files are typically ASCII text files in which there are either (i) a single column of stimulus regressor weights or (ii) a two-column format in which the first column indicates time into the functional run and the second column indicates the weight of the regressor. Matlab (v. 4.2 or greater).mat-file-based regressor files are also accepted.

Contributing authors are asked to provide the following in their preferred image data format:

- (i) Reconstructed images from the scanner.
- (ii) Preprocessed images (e.g. images that have gone through slice timing correction and/or spatial normalization. Putting the image data into this form is the stage just immediately prior to statistical analysis).
- (iii) Images of final statistical parametric map results (e.g. *t*-test images).
- (iv) All high-resolution anatomical images (e.g. those used for image registration, spatial normalization and the overlay of results).

In addition to the structural and functional MR data, the Data Center also encourages submission of supplemental data that is collected as part of the investigation, such as physiological recordings, behavioural data, etc. The Data Center aims to archive as much of the raw, preprocessed, anatomical and statistical image data as is possible in order to fully and completely represent each study, thereby permitting accurate reanalysis of the study data.

There are a few issues regarding quality control that are of vital importance in order to ensure the reliability of the information being stored. First, the Data Center only accepts data from peer-reviewed articles. Therefore, the quality control for these studies lies with the neuro-imaging community and its peer-review process. That is, since the study has been subjected to scholarly review by experts in the field and has been successfully published, it already has undergone the most rigorous degree of quality control assessment that could be expected. Data ancillary to the published study may be contributed but unless the authors provide a full description of that data and its necessity for understanding their reported results it will probably not be included with the data when distributed to other researchers. Unpublished data, although potentially valuable for offsetting publication bias, is likely to be more problematic than beneficial and will not be accepted. Second, all the information concerning a study is provided by the authors themselves and any imposition by the Data Center of presumed study information is strictly avoided. In no way does the Data Center wish to inadvertently alter or confuse reanalysis of the study data by altering or including variables not intended by the studies' authors. Third, the Data Center closely examines the agreement between the submitted data and that reported in the published article to check the integrity of the stored and disseminated data. Any inconsistencies are discussed and clarified with the contributing authors. Last, the final study description document that is constructed by the Data Center is sent to authors for their approval prior to making the study available to the public.

(b) Requesting study data

Requesting study data involves simply selecting the studies that the user would like and filling out a short Web form that sends an order notification to the Data Center. As of the beginning of 2001, all available study data may be obtained on either CD or DAT media. The study data is then shipped to requesting researchers in its

native format or converted to Mayo Analyse format for shipment. At present, no data are viewable online, although we anticipate that in future image data screening will be possible. In the future, we will also ship the data to users of the Data Center in the common medical image format of their choice and will offer users the ability to do basic meta-analyses of the studies stored within the fMRIDC database.

4. SUBJECT CONFIDENTIALITY AND CREDIT FOR ORIGINAL AUTHORS

Issues pertaining to protecting confidentiality are of fundamental importance in large neuroscientific databases that contain the raw experimental data from human research subjects. Additionally, the rights of the original authors of the published research article who deposit study data in the Data Center archive must be respected and duly considered in such efforts. These issues speak to the confidence that the research community has in database efforts in general, as well as the potential success of data-sharing initiatives, and deserve special attention.

(a) *The protection of human subject data*

The fMRIDC considers the protection of human subjects' data to be paramount to the success of the project. The Data Center makes every reasonable effort to ensure that only anonymized data are included in the Data Center archives and that researchers requesting data will only be using data that cannot be linked back to the individual subjects that provided it.

To comply with US Government regulations (45 CFR 46, referred to as 'The Common Rule') on the protection of human subject data, the Data Center asks the researchers submitting their study data to help in maintaining subject confidentiality. Here are the steps we advocate for authors submitting their functional imaging data to the Data Center:

- (i) Prior to sending data to the Data Center, submitting authors are asked to screen their data for subject identifiers such as name, subject initials, US Social Security Numbers, etc. as well as any internal subject identification codes used in their laboratory or at the submitting author's institution. These identifiers need to be removed altogether or changed such that there is no way in which the data can be linked back to the subject who provided it. This includes all image data files, image header files, behavioural data files, stimulus regressor files, etc.
- (ii) The Data Center realizes that some potential identifiers may have been missed in the screening done by the authors. In that light, once the Data Center has received the data, a further screening will be performed to ensure that all potential identifiers have been removed from the data. Again, this will encompass all image data files, image header files, behavioural data files, stimulus regressor files, etc. Compliance with US NIH Human Subjects requirements should be satisfied provided that every reasonable attempt has been made to remove identifiers, render anonymous and unlink the data by the authors submitting data as well as by the Data Center.

- (iii) As a particular concern to neuroimaging data, there is the possibility that surface reconstructions of high-resolution anatomical images could be used to identify subjects. To deal with this, the Data Center has adopted a simple approach: if researchers have a preferred method of stripping anatomical scans that they believe is essential for the accurate representation of their image data, then they are encouraged to submit those stripped anatomical image volumes to the Data Center along with the accompanying functional image data. If, however, they do not have a preferred method and simply submit the non-stripped original images then these will be stripped by the Data Center using the Statistical Parametric Mapping (Friston *et al.* 1995, 1996) software package or other suitable software. Only the stripped files will be made available to people who request study data.
- (iv) The Data Center encourages authors to download and read the informational memo concerning Human Subject Confidentiality (<http://www.fmridc.org/submission/humansubjects.php>). This may be submitted to their local Internal Review Board (IRB) to inform them of the Data Center's mission and the steps being taken to protect subject confidentiality. In addition, when submitting data to the Data Center, authors are encouraged to download and submit the IRB Study Submission Notification form (http://www.fmridc.org/documents/memo_06092000.php) to their local IRB. These procedures help to facilitate clear communication between the authors and their local IRBs on the compliance with the NIH requirements for the protection of human subjects.
- (v) For future studies that might be contributed to the Data Center, it is recommended that a statement be included in researcher's informed consent documentation that there is the potential that data collected from study participants may be made publicly available via the Data Center after it has been rendered anonymous. Based upon the feedback that the Data Center has received, there is no indication that this disclosure will affect people's willingness to participate in neuroimaging studies.

Again, it should be noted that these guidelines are implemented in order to satisfy US Government requirements and to ensure Data Center compliance with these legal regulations. The Data Center encourages researchers from around the world to contribute their neuroimaging data to the study archive but to do so only after careful consultation with their local institutional review committee and observing all governmental regulations of their home country for research involving human subjects.

(b) *The rights of original authors*

It is widely recognized that sometimes more information is collected in the process of a typical fMRI study than is originally reported. In this regard, in a letter to the US Office of Management and Budget, the President of the National Academy of Sciences stated that 'permitting the researcher who actually collected the data to be the first to analyse and publish conclusions concerning the data is an essential motivational aspect of research' (<http://www.nas.edu/includes/letter.htm>). The Data Center understands

this sentiment and appreciates the need for researchers to examine fully the effects in their data before sharing it with the research community at large. In order to allow authors more time to analyse their data, the Data Center will be following certain guidelines. The Data Center offers contributing authors the option of placing their data in a 'data hold' for a fixed period of time where visitors to the Data Center may view only the published statistical data and not the underlying functional and anatomical data. This basic procedure has already been implemented in other communities, namely the Protein Data Bank (<http://www.rcsb.org/pdb/nih-policy.html>). As the Data Center continues to evolve, these policies will continue to be reviewed and refined with regard to the best interests of the fMRI community.

Additionally, the Data Center encourages and expects researchers using data from the study archive, at the very least, to cite the work of the original authors, as well as the fMRIDC Accession no. of the study. This is standard practice in genomic database archives (for example, see The Stanford Malaria Genome Project, <http://sequence-www.stanford.edu/group/malaria/index.html>). To reanalyse data from the Data Center without at least citing the researchers responsible for generating the data we consider unacceptable and on a par with plagiarism. The Data Center is not in a position to have requirements over authorship agreements between the original authors and those wishing to publish reanalysed data. Researchers should be aware, however, of the potential for IRB considerations with such collaborations if there is the chance that the data may become relinked and therefore no longer anonymous.

5. THE fMRIDC AS A TOOL FOR PRIMARY RESEARCH

Accompanying the development of technologies for the construction of large, relational databases has been the increasing interest in methodologies for extracting and categorizing pertinent information from such repositories. Information retrieval in the medical sciences is enabling the rapid organization of information needed for both research and clinical purposes (Mavroudis & Jacobs 2000; Strasberg *et al.* 2000; Benoit & Andrews 2000; Wilcox & Hricpsak 1999; Goodman 1996). Additionally, data visualization and analysis methods have greatly improved with increases in the disk storage and memory capacity of desktop computers. Also, methods for the examination of results both within and between datasets has enabled researchers to examine the variables underlying the reported effects in published studies using the techniques of meta-analysis (for instance, see Van Horn & McManus 1992). Taken together, these areas form a new field of active research—that of neuroinformatics. The computational, analytic and visualization techniques being developed in this burgeoning area (see <http://www.nimh.nih.gov/neuroinformatics/index.cfm>) is an important and exciting aspect of the fMRIDC project.

(a) *The role of neuroinformatics*

Over the past decade, neuroinformatics research has begun to provide mechanisms for the storage, management and analysis of data from large neuroscientific data-

sets (Wong & Koslow 2001). The emerging discipline of neuroinformatics represents a bridging of the disciplines of computer science and informatics with that of neuroscience in order to use computational and mathematical methods to store, examine, categorize and model neuroscience data in order to extract pertinent information regarding brain structure and function (Huerta & Koslow 1996; Beltrame & Koslow 1999). In the USA, rapid developments in the neurosciences accompanied by decreases in the cost of storing data have hastened the need for developing database methodologies to manage the vast quantities of data that now exist (Bower 2000; Chicurel 2000). Databases containing data on large sets of neurons permit the analysis of cell morphology and computation properties (Jacobs 1996). Others enable the synthesis of information on the molecular interactions of neurons (Bloom 2000), as well as for searching metadata specifications across multidisciplinary databases (see for example, Shepherd *et al.* 1998). Principal among the challenges for neuroscientific databases is the enormous amount of data collected during *in vivo* studies of brain activity using fMRI.

A fundamental technique for the retrospective analysis of published data is that of meta-analysis (e.g. the 'analysis of analyses'). Using these techniques it is often possible to identify study factors that most influence effect-size estimates from statistical results across studies. In cognitive neuroscience, published neuroimaging results are now being quantitatively assessed to determine whether similar cognitive functions can be performed by disparate brain areas and whether individual brain areas are capable of performing disparate functions (for example, the implications for diseases like schizophrenia; see Zakzanis *et al.* (2000) and Zakzanis & Heinrichs (1999)). Typically, however, literature searches of heroic magnitude have been required by researchers to extract statistical values from published neuroimaging studies in order to conduct meta-analytic assessment of findings (e.g. see recently, Cabeza & Nyberg 2000). To facilitate meta-analytic assessment, pioneering work by Fox *et al.* (1998) to design and establish an online repository for the results of imaging experiments helped to advance the concepts of more dynamic meta-analytic assessment of findings from across numerous reports (BrainMap, <http://ric.uthscsa.edu/projects/brainmap.html>). However, for the neuroscientific benefit of fMRI for understanding brain structure–function relationships to be fully realized, comprehensive databases of complete, peer-reviewed, imaging studies of brain function, available to everyone, would be invaluable for enabling new and different perspectives on the interpretation of results (Young & Scannell 2000). The existence of such databases might spawn the need for methods of 'mega-analysis', that is, the statistical analysis of multiple, entire, raw datasets from across research centres to improve statistical power and provide even greater population generalizability.

(b) *Information retrieval at the fMRIDC*

Techniques from information retrieval will play a critical role in the utility and eventual widespread use of the fMRIDC archive. Presently, the search capabilities for the Data Center archive are relatively simple, being based on a MEDLINE-like search engine (figure 7).

The fMRI Data Center

[HOME](#) [ABOUT](#) [DATABASE](#) [SUBMISSIONS](#) [LINKS](#) [MY PROFILE](#)

Results Per Page:

[SETTINGS](#) [HISTORY](#) [CLIPBOARD](#) [LOAD](#)

Items 1-3 of 3 Page 1 of 1 One page.

Selected: Search Score: -Select Action-:


- 1: ☐ [Ishai A, Ungerleider LG, Martin A, Haxby JV](#)
The Representation of Objects in the Human Occipital and Temporal Cortex
 Journal of Cognitive Neuroscience 2000 Nov;12(suppl. 2):35-51
 FMRIDCID: 2-2000-1113D
- 2: ☐ [Arrington CM, Carr TH, Mayer AR, Rao SM](#)
Neural Mechanisms of Visual Attention: Object-Based Selection of a Region in Space
 Journal of Cognitive Neuroscience 2000 Nov;12(suppl. 2):106-117
 FMRIDCID: 2-2000-1116E
- 3: ☐ [Simpson JR, Ongur D, Akbudak E, Conturo TE, Ollinger JM, Snyder AZ, Gusnard DA, Raichle ME](#)
The Emotional Modulation of Cognitive Processing: An fMRI Study
 Journal of Cognitive Neuroscience 2000 Nov;12(suppl. 2):157-170
 FMRIDCID: 2-2000-1119F

Figure 7. Since many researchers are already familiar with the MEDLINE format, users may search the fMRIDC archive using a MEDLINE-inspired query interface. Searches may be done on the basis of author, author-supplied keywords and words in the abstract, among other possible variables. For instance, searching by the term 'occipital' returned the indicated list of studies. Search terms may be stored for future use in user profiles. The CD icon in the study citation indicates that the data from that published study is available for shipping.

However, this basic scheme has been modified to enable the weighting of search terms and the specification of loose (i.e. capable of handling simple misspelling of terms) or fuzzy (i.e. returning additional results based on terms similar to those entered by the user) searches (figure 8). However, even greater search capabilities are being planned for the future. Specifically, the application of information retrieval technology will further the transition of the Data Center from a mere warehouse of fMRI data to, in turn, (i) a dynamically searchable repository of fMRI studies, and (ii) a digital library of fMRI studies with user-driven and self-organizational capabilities. We discuss each of these phases in the subsections that follow.

For a large repository of data to be truly useful, one must be able to search efficiently for information of interest; techniques from the information retrieval community have been developed to address precisely this problem. Specifically, the stored data must be indexed to extract and store relevant features, and algorithms must be developed to match given queries to relevant indexed data.

Indexing is the process by which search features are extracted from data. For example, when processing a text document, the presence of interesting words (such as 'cerebellum' or 'visual') might correspond to search features; when processing statistical maps, high z -scores

The fMRI Data Center

[HOME](#) [ABOUT](#) [DATABASE](#) [SUBMISSIONS](#) [LINKS](#) [MY PROFILE](#)

Results Per Page:

[SETTINGS](#) [HISTORY](#) [CLIPBOARD](#) [LOAD](#)

Matching Level

		Exact	Loose	Fuzzy	Weight
[AU]	Author	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="text" value="1.0"/>
[AB]	Abstract	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="text" value="1.0"/>
[FMRIDCID]	Accession#	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text" value="1.0"/>
[KW]	Article Keywords	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="text" value="1.0"/>
[PMID]	PubMed Id	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text" value="1.0"/>
[TI]	Title	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="text" value="1.0"/>

Publication Date From to
Date format: YYYY/MM/DD. Month and day are optional.

Figure 8. Currently, a key feature of the fMRIDC database search capability is that the different search fields can be modified to the weighting given to them. They may be further constrained to be 'exact', 'loose' or 'fuzzy' to alter the degree of match desired. 'Exact' refers to a precise match of a search term (including case); a 'loose' search allows for simple misspellings (e.g. the inversion of characters) and would be case insensitive; and 'fuzzy' refers to searches that not only look for the entered search terms but also include results from similar or related terms. Search terms can also be extended using Boolean operators. The results of searches are stored for the user so that they can recall search terms as new datasets are added to the database. In the future, the Data Center's query interface will be expanded to include greater complexity of search schemes and include sophisticated data-mining capabilities.

present at a particular Talairach & Tournoux (1988) coordinate might correspond to search features. For the purpose of conducting a search, data objects may then be represented by their basic search features (e.g. Document A contains the interesting words 'cerebellum', 'visual' and 'stimuli'; Statistical Map B contains high z -scores in the cerebellum, etc.). Note that the total size of the search features may be orders of magnitude smaller than the total size of their corresponding data objects; thus, they may be stored in an online database and queried efficiently.

The search features obtained through indexing are a type of 'metadata'; i.e. information stored about the data present in the repository. Other types of metadata may be stored as well. The existence of common metadata corresponding to all the data stored in a repository permits the uniform application of search (and higher level information retrieval techniques) across that data. Similarly, the existence of common metadata across repositories would in principle permit the uniform application of searches across repositories. For instance, valuable work done on probabilistic brain atlases by the International Consortium for Brain Mapping (ICBM) (<http://www.loni.ucla.edu/ICBM/index.html>) has established reference frames for population-level average and variance estimates for normal brain structure that may be used to evaluate neurological patients (Mazziotta *et al.* 1995; Mazziotta 2000; Thompson *et al.* 2000). For instance, a

clinician or investigator who obtains an MRI scan from a patient having focal epilepsy can call on the ICBM probabilistic brain atlas of healthy subjects and compare the patient's brain with the average normal brain. Once the clinician has identified that the structure of a particular brain location in this patient is at variance with what is expected based upon the probabilistic atlas, he or she might then wish to query other neuroscientific databases. Metadata obtained from the ICBM atlas comparison might then be passed into a search query at the fMRIDC in order to identify functional imaging studies in which activation of that brain area was reported. Metadata from both searches may then be submitted to another search of a database of neuron cell types (for instance, the Yale SENSELAB database (Marenco *et al.* 1999, <http://ycmi-hbp.med.yale.edu/senselab/>) to find information on the principal cells residing in that area, their neurotransmitters, etc. Finally, all this information could be hierarchically formatted and a report presented to the clinician giving a complete picture of the current understanding of the affected brain area in this patient. This may, in turn, lead to a more targeted course of treatment and a better understanding of the concomitants of lesions to that area. Standardized metadata frameworks (based on extensible mark-up language (XML), for instance) that lend themselves to efficient information management and retrieval from neuroscientific databases will facilitate the rapid occurrence of such scenarios in the not too distant future.

Given a database of search features ('metadata') corresponding to fMRI studies, query retrieval algorithms can be applied to retrieve whole fMRI studies (or related individual data objects) relevant to a given search query. The queries themselves may simply be keywords, Boolean combinations thereof, or more sophisticated specifications involving ranges, proximity, fuzzy search, etc. Our expectation is that the current MEDLINE-like query interface will evolve over time as more and interesting features are indexed and stored in the database. The search algorithm itself is a procedure that takes a query and database of search features as input and returns a (typically ranked) collection of relevant data objects as output. Many search algorithms have been developed in the information retrieval community based on various mathematical formalisms, including the vector space model, Bayesian networks, the probabilistic inference model, etc. (Baeza-Yates & Ribeiro-Neto 1999; Croft 2000; Jones & Willett 1997). Any of these search algorithms can be used at the Data Center, and transitioning from one to another would be seamless to an end user. The vector space model holds particular promise for both search and automatic organization, as will be outlined in § 5(c). A given search at the Data Center could be processed locally, and it could also be automatically sent to search engines at related sites such as PubMed, etc. The results from these multiple searches (both local and non-local) could then be intelligently combined in a single response for the end user, thus harnessing the power of multiple search engines over disparate repositories of related information. We have developed a number of 'metasearch' algorithms for just this purpose and propose to deploy them for use by the neuroscience community (Aslam & Montague 2000, 2001).

(c) *The application of information retrieval techniques for automatic organization*

Sophisticated search, the ability to efficiently find relevant pieces of information within a collection of data, distinguishes a digital library from a mere repository of information, as described in the preceding section. The next phase in the evolution of the Data Center will be the ability to automatically organize information subject to user-specified criteria. Text documents, for example, could be automatically organized into groups of similar documents via content analysis (e.g. documents containing identical or similar keywords would in all likelihood be about similar topics and could be grouped together). Statistical maps might be judged similar and grouped together if they had similar patterns of activation according to a given metric. Whole fMRI studies could be judged similar and grouped together if they had similar experimental protocols and empirical findings.

In order to automatically organize information, a similarity metric might be defined, which assesses the similarity of two data objects, and then a clustering algorithm may be used, which, given the pairwise similarities of a collection of data objects, partitions these objects into self-similar groups. Similarity metrics are generally data-specific. For example, the similarity between two statistical maps could be found by computing the root mean squared difference between the *z*-scores at selected Talairach-Tournoux coordinates. A generic technique also exists which is often applied in the information retrieval community. Data objects are represented by their features, as described in the preceding section, and can be considered as vectors in 'feature space'. For example, the relevant features of a text abstract might be the presence or absence of various keywords. An abstract can then be represented by vector in 'keyword space', where the value associated with each coordinate is 1 or 0 if the corresponding keyword is present or absent, respectively, in the abstract. The similarity of two abstracts can then be assessed by determining how 'close' the corresponding vectors are to each other in keyword space (typically, by computing the cosine of the angle between the vectors). This technique can be applied generically to data objects that can be represented as numerical feature vectors. We have developed a number of generic similarity metrics (geometric, probabilistic and information-theoretic), and we are developing, in an on-going manner, specific similarity metrics for neuroscientific data.

Given the ability to assess the pairwise similarity of two data objects, clustering techniques can be employed to partition the objects into self-similar groups. Many clustering algorithms have been developed within and outside the information retrieval community (Kaufman & Rousseeuw 1990), and we have developed a number of efficient and accurate clustering algorithms in-house as well (Aslam *et al.* 1998*a,b*, 1999, 2000). We envisage using these automatic organization tools in at least two ways. As a preprocessor, our tools can be used to organize all the studies stored at the Data Center according to user-specified criteria. For example, a user may wish to browse the studies stored at the Data Center organized according to author, research laboratory, experimental protocol, statistical findings (activation sites) or some combination thereof. The user could simply specify the organizational

criteria, and our tools can then automatically provide a graphical user interface to the data organized as requested (we have developed a number of more sophisticated user-interfaces as well). As a postprocessor, such tools can be used to automatically organize the results of a search. For example, a user could request to browse all studies at the Data Center that are event-related and involve visual stimuli, organized according to similarity of activation patterns. Finally, these tools are dynamic in the sense that a given organization need not be entirely recomputed as the underlying data changes (i.e. as more studies are added to the Data Center); new data can be added to an existing organization immediately. A user's 'view' of the data (his or her 'profile') can be saved and updated efficiently when the user returns. Furthermore, this permits the use of user-specified persistent queries and automatic notification of interesting new studies. For example, the user, having found one or more collections of studies of particular interest, may 'flag' those collections for automatic notification. As new studies are added to the Data Center archive, they can automatically and efficiently be added to the stored user-specified organization as well. If a new study is added to a flagged collection, the user can be automatically notified of a new study of interest via email. We have already developed and deployed much of the technology described for digital libraries consisting of textual data. Since our organization tools rely on similarity scores alone and not specific properties of text documents, they can be employed to organize neuroscientific data as well, given appropriate similarity metrics. We are in the process of developing such metrics and deploying our automatic organization tools for the Data Center.

6. FUTURE CHALLENGES

The Data Center expects to advance its research component by designing and building useful tools for neuroimaging that help researchers search, categorize, and obtain imaging study data. Additionally, it is hoped that the educational component of the Data Center will be realized and the archive may provide a valuable instructional resource in a manner similar to the National Library of Medicine's Visible Human Project (http://www.nlm.nih.gov/research/visible/visible_human.html), where students and researchers can download images to investigate human anatomy (for instance, see Schiemann *et al.* 2000). To do this the Data Center needs to overcome a number of significant technical and intellectual hurdles. We note two of these challenges here.

(a) *Metadata standards for fMRI experiments*

Equally important as the overall design of the host database is being able to represent these data in such a fashion that it permits the essential information about studies to be distilled into a manageable form which then allows study data to be efficiently searched, organized and categorized. This will require that the architecture of the metadata schema be self-describing (all the information needed to 'describe the description' of the data will be encapsulated in the metadata framework), expandable (capable of incorporating exemplars for different neuroscience variables), and extensible (capable of

growing as the field grows to include more pertinent information about the research being conducted). Given the aforementioned challenges of describing neuroscientific experimentation, the development of such metadata architecture will need to be carefully considered (e.g. Gardner *et al.* 2001). Not only should such a format satisfy the needs of functional neuroimaging experiments (and, thus, useful for the fMRIDC), it should be sufficiently generalizable to be used for other forms of neuroscientific research. It should also be sensitive to the importance of human subjects research and not include information that can be linked back to individual subjects. Moreover, it should facilitate interoperability between other database architectures so as to enable researchers to identify lines of converging evidence that explain neuroscientific findings. Finally, by having advanced search capability through a fully extensible framework, this will enable the fMRIDC to increase its user base, and, thus, speed progress in the field of cognitive neuroscience.

(b) *Efficient image file compression schemes for fMRI time-series*

Image file compression is an area where the expertise of mathematicians and computer scientists can make substantial contributions to fMRI data management. For instance, using wavelet-based compression schemes, it is possible to reduce the amount of redundant information contained in fMRI time-series and store the data in much less space than the original time-series (Chawla 1998). Additional techniques that can be drawn upon include Karhunen–L   ve (K-L) decompositions (Wickerhauser 1994; Weaver & Healy 1994), robust wavelet approximations to the K-L decomposition (Healy & Weaver 1996; Healy *et al.* 1997; Olson *et al.* 1994), and multifocal decompositions that classify the images into groups where different compression methods are used (Chawla 1998). These and other methods need to be rigorously examined for their suitability, robustness and reliability for fMRI data where investigators will probably not tolerate loss processes that might adversely affect the detection of cognitive activation. However, these and similar methods may permit more rapid and efficient transfer of data over the Internet and, thus, make downloading study data from the Data Center easier.

(c) *Advancing the educational component of the fMRIDC*

Centrally important is the educational role that the Data Center may play in the training of researchers in methodologies that would permit novel examination of fMRI data as well as in developing new tools that might be useful to the neuroimaging community as a whole. Bioinformatics educational programmes have been established both in the United States (Altman 1998) and in Europe (Brass 2000), that have tended to focus on undergraduate training on medical, biological and genomic informatics. Research-orientated programmes in neuroinformatics may further help to expedite progress in this area by speeding the development of tools and algorithms that may help researchers in understanding their fMRI datasets. Bringing engineers, mathematicians, statisticians and neuroscientists who already have advanced degrees in their own fields into the realm of

neuroinformatics would help to increase the number of practising neuroinformaticians more rapidly than by waiting until they have completed undergraduate training and then undertaken research careers. To do this, however, it will be necessary to encourage both computer scientists and biomedical engineering specialists to take on the special informatics challenges that confront the field of fMRI and which will benefit neuroscience as a whole. This will require some additional training that will help to familiarize researchers from outside neuroscience with the fundamentals of brain research and to introduce concepts in neuroscience to computer scientists, thus providing the background needed to then take on the computational challenges in an informed manner.

7. CONCLUSION

By starting from the simple but explicit goal of facilitating sharing of published fMRI studies, the fMRIDC hopes to advance the progress being currently made in understanding cognitive function. As we point out above, several challenges still need to be overcome before the fMRIDC reaches its ambitious goals. These challenges are currently areas of active research and study. Yet, once surmounted, there is the potential for efforts like the fMRIDC, acting as an official record for the scientific body of work in the area of neuroimaging studies, to affect a paradigm shift in the current thinking of what it means to have a study published. For a researcher to have his/her published results, and their interpretation, accepted by the neuroscientific community, it may be expected that their raw data be deposited in a published study archive like the fMRIDC. This sentiment is embodied in calls for a greater amount of data sharing among brain scientists and the development of additional database projects for the neurosciences (Koslow 2000; Editorial 2000). In future, therefore, contribution of data to a neuroscientific data archive will have simply become an expected part of the peer-reviewed publication process.

The fMRIDC effort aims to maintain the scientific record of published functional neuroimaging studies submitted to its archive and to develop the tools needed to develop fully this portion of the data-sharing continuum. In so doing, the Data Center hopes to be able to facilitate the greater involvement of scientists from a diverse range of backgrounds, bringing to bear their expertise in solving the riddles of how the brain works. The ultimate rewards of this endeavour will be realized when more studies are examined together than when any single study is examined alone.

This work is funded by a grant from the US National Science Foundation (Award BCS-9978116), The National Institute of Mental Health and generous support from the W. M. Keck Foundation (Award 991714). The Data Center also receives support from Informix Software, Inc. through an Academic Research Innovation Grant for software education and training using the Informix database platform. The authors wish to extend their thanks to Dr Souheil Inati and Dr Mark Wessinger, Mr Mark Montague, Mr Derek Nee, Ms Luminita Dirna and Ms Wendy Starr of the fMRIDC for their valuable contributions to this project.

REFERENCES

- Altman, R. B. 1998 A curriculum for bioinformatics: the time is ripe. *Bioinformatics* **14**, 549–550.
- Arbib, M. A. & Grethe, J. S. (eds) with the Project Team of the University of Southern California Brain Project 2001 *Computing the brain: a guide to neuroinformatics*. San Diego, CA: Academic Press.
- Arbib, M. A., Grethe, J. S., Wehrer, G. L., Mureika, J. R., Tracy, J., Xie, X., Thompson, R. F. & Berger, T. W. 1996 An on-line neurophysiological and behavioral database for the neuroscientist. *Soc. Neurosci. Abstr.* **2**, 359.16.
- Aslam, J. & Montague, M. 2000 Bayes optimal metasearch: a probabilistic model for combining the results of multiple retrieval systems. In *Proceedings of the Twenty-Third International Association for Computing Machinery Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval* Athens, Greece, 24–28 July 2000.
- Aslam, J. & Montague, M. 2001 Models for metasearch. In *Association for Computing Machinery Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval* New Orleans, LA, USA, 9–13 September 2001.
- Aslam, J., Pelekhev, K. & Rus, D. 1998a Static and dynamic information organization with star clusters. In *Proceedings of the 1998 International Conference on Information and Knowledge Management*, Bethesda, MD, USA, 3–7 November 1998.
- Aslam, J., Pelekhev, K. & Rus, D. 1998b Generating, visualizing and evaluating high-quality clusters for information organization. In *Principles of digital document processing: 4th International Workshop* (ed. E. V. Munson, C. Nicholas & D. Wood), pp. 53–69. New York: Springer.
- Aslam, J., Pelekhev, K. & Rus, D. 1999 A practical clustering algorithm for static and dynamic information organization. In *Proceedings of the Tenth Association for Computing Machinery, Society for Industrial and Applied Mathematics Symposium on Discrete Algorithms*, Baltimore, MD, USA, 17–19 January 1999.
- Aslam, J., Pelekhev, K. & Rus, D. 2000 Using star clusters for filtering. In *Proceedings of the 2000 Association for Computing Machinery International Conference on Information and Knowledge Management*.
- Bacza-Yates, R. & Ribeiro-Neto, B. 1999 *Modern information retrieval*. New York: ACM Press.
- Beltrame, F. & Koslow, S. H. 1999 Neuroinformatics as a megascience issue. *Institute for Electrical and Electronics Engineers Trans. Inform. Technol. Biomed.* **3**, 239–240.
- Benoit, G. & Andrews, J. E. 2000 Data discretization for novel resource discovery in large medical data sets. In *Proceedings of the American Medical Informatics Association Symposium*, pp. 61–65.
- Bloom, F. E. 1996 The multidimensional database and neuroinformatics requirements for molecular and cellular neuroscience. *NeuroImage* **4**, S12–S13.
- Bower, J. M. 2000 What will save neuroscience? *NeuroImage* **4**, S29–S33.
- Brass, A. 2000 Bioinformatics education—a UK perspective. *Bioinformatics* **16**, 77–78.
- Cabeza, R. & Nyberg, L. 2000 Imaging cognition II: an empirical review of 275 PET and fMRI studies. *J. Cogn. Neurosci.* **12**, 1–47.
- Casey, B. J. (and 11 others) 1998 Reproducibility of fMRI results across four institutions using a spatial working memory task. *NeuroImage* **8**, 249–261.
- Chawla, S. 1998 Novel compression techniques with applications in medical image acquisition and image storage. PhD thesis, Dartmouth College, Hanover, NH.
- Chen, I. A. & Markowitz, V. M. 1995 An overview of the object-protocol model (OPM) and the OPM data management tools. *Inform. Sys.* **20**, 393–417.

- Chen, P. P. 1976 The entity-relationship model—toward a unified view of data. *Association for Computing Machinery Trans. Database Sys.* **1**, 9–36.
- Chicurel, M. 2000 Databasing the brain. *Nature* **406**, 822–825.
- Cox, R. W. 1996 AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* **29**, 162–173.
- Croft, W. B. 2000 *Advances in information retrieval: recent research from the center for intelligent information retrieval*. New York: Kluwer.
- Dalton, R. 2000 Young worldly and unhelpful all miss out on data sharing. *Nature* **404**, 6.
- Editorial 2000 A debate over fMRI data sharing. *Nat. Neurosci.* **3**, 845–846.
- Fox, P. T., Parsons, L. M. & Lancaster, J. L. 1998 Beyond the single study: function/location metanalysis in cognitive neuroimaging. *Curr. Opin. Neurobiol.* **8**, 178–187.
- Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D. & Frackowiak, R. S. J. 1995 Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* **2**, 189–210.
- Friston, K. J., Poline, J.-B., Holmes, A. P., Frith, C. D. & Frackowiak, R. S. J. 1996 A multivariate analysis of PET activation studies. *Hum. Brain Mapp.* **4**, 140–151.
- Gardner, D., Knuth K. H., Abato, M., Erde, S. M., White, T., DeBellis, R. & Gardner E. P. 2001 Common data model for neuroscience data and data model exchange. *J. Am. Med. Inform. Assoc.* **8**, 17–33.
- Goodman, K. W. 1996 Ethics, genomics and information retrieval. *Comput. Biol. Med.* **26**, 223–229.
- Grethe, J. S. 1999 Neuroinformatics and the cerebellum: towards an understanding of the cerebellar microzone and its contribution to the well-timed classically conditioned eyeblink response. PhD thesis, University of Southern California.
- Grethe, J. S. & Grafton, S. T. 1999 An on-line experimental database for the storage and retrieval of neuroimaging data. *Soc. Neurosci. Abstr.* **25**, 4.52.
- Grethe, J. S., Wehrer, G. L., Thompson, R. F., Berger, T. W. & Arbib, M. A. 1996 An extendible object-relational database schema for neurophysiological and behavioral data. *Soc. Neurosci. Abstr.* **22**, 359.17.
- Grethe, J. S., Mureika, J. & Merchant, E. N. 2001 Design concepts for a neuroscience database. In *Computing the brain: a guide to neuroinformatics* (ed. M. A. Arbib & J. S. Grethe), pp. 135–150. San Diego, CA: Academic Press.
- Healy, D. M. & Weaver, J. B. 1996 Adapted wavelet techniques for encoding magnetic resonance images. In *Wavelets in medicine and biology* (ed. A. Aldroubi & M. Unser). CRC Press, Boca Raton, FL, USA.
- Healy, D. M., Warner, D., Chawla, S. & Weaver, J. B. 1997 *Adapted waveform encoding and fast magnetic resonance imaging*. The first international congress of The International Society for Analysis, its Applications and Computation (ISAAC), 3–7 June 1997, University of Delaware, Newark, DE.
- Huerta, M. F. & Koslow, S. H. 1996 Neuroinformatics: opportunities across disciplinary and national borders. *NeuroImage* **4**, S4–S6.
- Jacobs, G. A. 1996 Analysis of information processing in the nervous system using a database of identified neurons. *NeuroImage* **4**, S23–S24.
- Jones, K. S. & Willett, P. 1997 *Readings in information retrieval*. San Francisco, CA: Morgan Kaufmann.
- Kastner, T. 2000 On the need for policy requiring data-sharing among researchers publishing in AAMR journals: critique of Conroy and Adler 1998. *Men. Retard.* **38**, 519–529.
- Kaufman, L. & Rousseeuw, P. J. 1990 *Finding groups in data: an introduction to cluster analysis*. New York: Wiley.
- Koslow, S. H. 2000 Should the neuroscience community make a paradigm shift to sharing primary data? *Nat. Neurosci.* **3**, 863–865.
- Kostelec, P., Grethe, J., Aslam, J., Rockmore, D., Fendrich, R., Grafton, S. & Gazzaniga, M. S. 2000 A National Data Center for the storage and retrieval of neuroimaging data. *Soc. Neurosci. Abstr.* **26**, 2235(840.7).
- Marenco, L., Nadkarni, P., Skoufos, E., Shepherd, G. & Miller, P. 1999 Neuronal database integration: the Senselab EAV data model. In *Proceedings of the American Medical Informatics Association Symposium*, pp. 102–106.
- Mavroudis, C. & Jacobs, J. P. 2000 Congenital heart surgery nomenclature and database project: overview and minimum dataset. *Annls Thorac. Surg.* **69**(Suppl. 4), S2–S17.
- Mazziotta, J. C. 2000 Imaging: window on the brain. *Arch. Neurol.* **57**, 1413–1421.
- Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P. T. & Lancaster, J. 1995 A probabilistic atlas of the human brain: theory and rationale for its development. *NeuroImage* **2**, 89–101.
- Mechelli, A., Friston, K. J. & Price, C. J. 2000 The effects of presentation rate during word and pseudoword reading: a comparison of PET and fMRI. *J. Cogn. Neurosci.* **12**(Suppl. 2), 145–156.
- Olson, T., Healy, D. M. & Weaver, J. B. 1994 Reconstruction via adaptive spatio-temporal acquisition for motion induced noise. *Proceedings of the International Society for Optical Engineering*.
- Opinion 2000 Whose scans are they, anyway? *Nature* **406**, 443.
- Postle, B. R., Berger, J. S., Taich, A. M. & D'Esposito, M. 2000 Activity in human frontal cortex associated with spatial working memory and saccadic behavior. *J. Cogn. Neurosci.* **12**(Suppl. 2), 2–14.
- Schiemann, T., Freudenberg, J., Pflesser, B., Pommert, A., Priesmeyer, K., Riemer, M., Schubert, R., Tiede, U. & Hohne, K. H. 2000 Exploring the visible human using the VOXEL-MAN framework. *Comput. Med. Imaging Graph.* **24**, 127–132.
- Shepherd, G. M., Mirsky, J. S., Healy, M. D., Singer, M. S., Skoufos, E., Hines, M. S., Nadkarni, P. M. & Miller, P. L. 1998 The Human Brain Project: neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data. *Trends Neurosci.* **21**, 460–468.
- Strasberg, H. R., Manning, C. D., Rindfleisch, T. C. & Melmon, K. L. 2000 What's related? Generalizing approaches to related articles in medicine. In *Proceedings of the American Medical Informatics Association Symposium*, pp. 838–842.
- Talairach, J. & Tournoux, P. 1988 *Co-planar stereotaxic atlas of the human brain*. Stuttgart, Germany: Thieme.
- Thompson, P. M., Woods, R. P., Mega, M. S. & Toga, A. W. 2000 Mathematical/computational challenges in creating deformable and probabilistic atlases of the human brain. *Hum. Brain Mapp.* **9**, 81–92.
- Thompson, R. F., Grethe, J. S., Berger, T. W. & Xie, X. 2001 Repositories for the storage of experimental neuroscience data. In *Computing the brain: a guide to neuroinformatics*, (ed. M. A. Arbib & J. S. Grethe), pp. 117–133. San Diego, CA: Academic Press.
- Van Horn, J. D. & McManus, I. C. 1992 Ventricular enlargement in schizophrenia. A meta-analysis of studies of the ventricle:brain ratio (VBR). *Br. J. Psychiat.* **160**, 687–697.
- Weaver, J. B. & Healy, D. M. 1994 Acquisition of the Karhunen–L  eve expansion to reduce MR imaging times. In *Institute for Electrical and Electronics Engineers International Conference on Image Processing, Austin, Texas 1994*, vol. III, p. 35.
- Wickerhauser, M. V. 1994 *Adaptive wavelet analysis from theory to software*. Wellesley: A. K. Peters Ltd., Natick, MA, USA.
- Wilcox, A. & Hripcsak, G. 1999 Classification algorithms applied to narrative reports. *Proceedings of the American Medical Informatics Association Symposium*, pp. 455–459.

- Wong, S. T. & Koslow, S. H. 2001 Human brain program research progress in bioinformatics/ neuroinformatics. *J. Am. Med. Inform. Assoc.* **8**, 103–104.
- Young, M. P. & Scannell, J. W. 2000 Brain structure–function relationships: advances from neuroinformatics. *Phil. Trans. R. Soc. London B* **355**, 3–6.
- Zakzanis, K. K. & Heinrichs, R. W. 1999 Schizophrenia and the frontal brain: a quantitative review. *J. Int. Neuropsychol. Soc.* **5**, 556–566.
- Zakzanis, K. K., Poulin, P., Hansen, K. T. & Jolic, D. 2000 Searching the schizophrenic brain for temporal lobe deficits: a systematic review and meta-analysis. *Psychol. Med.* **30**, 491–504.